In the Specification

Please amend the specification on page 6 line 25 to page 7 line 7 as follows:

Similarity analysis includes database search and alignment. Examples of public databases include the DNA Database of Japan (DDBJ) ~~(http://www.ddbj.nig.ac.jp/)~~ (www-ddbj.nig.ac.jp/); Genebank ~~(http://www.ncbi.nlm.nih.gov/web/Genbank/Index.htlm)~~ (www-ncbi.nlm.nih.gov/web/Genbank/Index.html); and the European Molecular Biology Laboratory Nucleic Acid Sequence Dabatase (EMBL) ~~(http://www.ebi.ac.uk/ebi_docs/embl_db.html)~~ www-ebi.ac.uk/ebi_docs/embl_db.html). A number of different search algorithms have been developed, one example of which are the suite of programs referred to as BLAST programs. There are five implementations of BLAST, three designed for nucleotide ~~sequences~~ sequence queries (BLASTN, BLASTX, and TBLASTX) and two designed for protein sequence queries (BLASTP and TBLASTN) (Coulson, *Trends in Biotechnology*, *12*:76-80 (1994); Birren, *et al.*, *Genome Analysis*, *1*:543-559 (1997)).

Please amend the specification on page 11 lines 3 to 15 as follows:

A characteristic feature of a large scale shotgun sequencing project is that the sequence data can be processed and assembled into contiguous sequences (contigs), which represent a reconstruction of the original genome sequence from the cloned fragments. Likewise, individual Bacterial Artificial Chromosome (BAC) clones within a BAC library can be shot gun sequenced and these data can be assembled into contigs. Programs are available in the public domain that can analyze the sequence output and assemble the sequences into larger sequence regions

representing contiguous sequences of the target genome. Examples of such programs can be found at, for example, ~~http://genome.wustl.edu/gsc~~ <u>genome.wustl.edu/gsc</u>, ~~http://www.sanger.ac.uk~~ <u>www-sanger.ac.uk</u>, and ~~http://www.mbt.washington.edu~~ <u>www-mbt.washington.edu</u>. An example of <u>a</u> sequence reading program is ~~Phred~~ <u>PHRED</u> ~~(http://www.mbt.washington.edu)~~ <u>(www-mbt.washington.edu)</u>. ~~Phred~~ <u>PHRED</u> reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files.

Please amend the specification on page 11 line 16 to page 12 line 6 as follows:

The process of assembling DNA sequence fragments generally involves three phases; the overlap phase, the layout phase and the multi-alignment, or consensus, phase. In the overlap phase, each fragment is compared against every other fragment to determine if they share a common subsequence, an indication that they were potentially sampled from overlapping stretches of the original DNA strand. Pairs of fragments are compared in two ways; 1) with both fragments in the same relative orientation, and 2) with one of the fragments having been reverse complemented. In the layout phase, a series of alternate assemblies or layouts of the fragments based on the pairwise overlaps is generated. A layout specifies the relative locations and orientations of the fragments with respect to each other and is typically visualized as an arrangement of overlapping directed lines, one for each fragment. The general criterion for the layout phase is to produce plausible assemblies of maximum likelihood. In this manner, it can be determined if there is more than one way to put the pieces together and if different solutions appear equally plausible. The multi-alignment, or consensus, phase uses more information than

just the pairwise alignments in the layout. The sequences of all the fragments in a layout are simultaneously aligned, giving a final set of contigs representing regions of the target genome. An example of an assembly program is PHRAP, which can be found at ~~http://chimera.biotech.washington.edu/UWGC/tools/phrap.htm~~ chimera.biotech.washington.edu/UWGC/tools/phrap.htm.

Please amend the specification on page 15 line 25 to page 16 line 2 as follows:

The present invention also provides a substantially purified nucleic acid molecule encoding a rice protein or fragment thereof, wherein the rice protein or fragment thereof is encoded by a nucleic acid sequence selected from the group consisting of [[52202]] SEQ ID NO: 1 through SEQ ID NO: 52202 or complements thereof or fragments of either.

Please amend the specification on page 25 line 24 to page 26 line 25 as follows:

SNPs can be characterized using any of a variety of methods. Such methods include the direct or indirect sequencing of the site, the use of restriction enzymes (Botstein *et al.*, *Am. J. Hum. Genet. 32*:314-331 (1980), the entirety of which is herein incorporated by reference; Konieczny and Ausubel, *Plant J. 4*:403-410 (1993), the entirety of which is herein incorporated by reference), enzymatic and chemical mismatch assays (Myers *et al.*, *Nature 313*:495-498 (1985), the entirety of which is herein incorporated by reference), allele-specific PCR (Newton *et al.*, *Nucl. Acids Res. 17*:2503-2516 (1989), the entirety of which is herein incorporated by reference; Wu *et al.*, *Proc. Natl. Acad. Sci. USA 86*:2757-2760 (1989), the entirety of which is herein incorporated by reference), ligase chain reaction (Barany, *Proc. Natl. Acad. Sci. USA*

88:189-193 (1991), the entirety of which is herein incorporated by reference), single-strand conformation polymorphism analysis (Labrune *et al.*, *Am. J. Hum. Genet. 48*: 1115-1120 (1991), the entirety of which is herein incorporated by reference), primer-directed nucleotide incorporation assays (Kuppuswami *et al.*, *Proc. Natl. Acad. Sci. USA 88*:1143-1147 (1991), the entirety of which is herein incorporated by reference), dideoxy fingerprinting (Sarkar *et al.*, *Genomics 13*:441-443 (1992), the entirety of which is herein incorporated by reference), solid-phase ELISA-based oligonucleotide ligation assays (Nikiforov *et al.*, *Nucl. Acids Res. 22*:4167-4175 (1994), the entirety of which is herein incorporated by reference), oligonucleotide fluorescence-quenching assays (Livak *et al.*, *PCR Methods Appl. 4*:357-362 (1995[[a]]), the entirety of which is herein incorporated by reference), 5'-nuclease allele-specific hybridization ~~TaqMan~~™ <u>TAQMAN™</u> assay (Livak *et al.*, *Nature Genet. 9*:341-342 (1995), the entirety of which is herein incorporated by reference), template-directed dye-terminator incorporation (TDI) assay (Chen and Kwok, *Nucl. Acids Res. 25*:347-353 (1997) the entirety of which is herein incorporated by reference), allele-specific molecular beacon assay (Tyagi *et al.*, *Nature Biotech. 16*: 49-53 (1998), the entirety of which is herein incorporated by reference), ~~PinPoint~~ <u>PINPOINT™</u> assay ([[ ]]Haff and Smirnov, *Genome Res. 7*: 378-388 (1997), the entirety of which is herein incorporated by reference), and dCAPS analysis (Neff *et al.*, *Plant J. 14*:387-392 (1998), the entirety of which is herein incorporated by reference).

Please amend the specification on page 37 lines 11 to 18 as follows:

Genomic sequences can be screened for the presence of protein homologues or genes utilizing one or a number of different search algorithms ~~have~~ that <u>have</u> been developed, one

example of which are the suite of programs referred to as BLAST programs. Other examples of suitable programs that can be utilized are known in the art, several of which are described above in the Background and under the section titled "Exemplary Uses of the Agents of the Invention." In addition, unidentified reading frames may be screened for protein coding regions by prediction sofware such as ~~GenScan~~ GENSCAN™, which is located at ~~http://genomic.standford.edu/GENSCANW.html~~ gnomic.stanford.edu/GENSCANW.html.

Please amend the specification on page 42 lines 9 to 15 as follows:

Nucleic acid molecules of the present invention can comprise an intron and/or one or more intron/exon ~~junction~~ junctions. Sequences of the present invention can be screened for introns and intron/exon junctions utilizing one or a number of different search algorithms that have ~~that~~ been developed, one example of which are the suite of programs referred to as BLAST programs. Other examples of suitable programs that can be utilized are known in the art, several of which are described above in the Background and in the section entitled "Exemplary Uses of the Agents of the ~~Present~~ Invention."

Please amend the specification on page 45 lines 16 to 26 as follows:

Murine monoclonal antibodies are particularly preferred. BALB/c mice are preferred for this purpose, however, equivalent strains may also be used. The animals are preferably immunized with approximately 25 μg of purified protein (or fragment thereof) that has been emulsified in a suitable adjuvant (such as ~~TiterMax~~ TITERMAX® adjuvant (Vaxcel, Norcross, GA)). Immunization is preferably conducted at two intramuscular sites, one intraperitoneal site,

and one subcutaneous site at the base of the tail. An additional i.v. injection of approximately 25 μg of antigen is preferably given in normal saline three weeks later. After approximately 11 days following the second injection, the mice may be bled and the blood screened for the presence of anti-protein or peptide antibodies. Preferably, a direct binding Enzyme-Linked Immunoassay (ELISA) is employed for this purpose.

Please amend the specification on page 74 lines 1 to 10 as follows:

Exogenous genetic material may be transferred into a plant cell by the use of a DNA vector or construct designed for such a purpose. Preferred exogenous genetic material comprise a nucleic acid molecule of the present invention. Vectors have been engineered for transformation of large DNA inserts into plant genomes. Vectors have been designed to replicate in both *E. coli* and *A. tumefaciens* and have all of the features required for transferring large inserts of DNA into plant chromosomes (Choi and Wing, ~~http://genome.clemson.edu/protocols2-nj.html~~ <u>genome.clemson.edu/protocols2-nj.html</u> July, 1998). ApBACwich system has been developed to achieve site-directed integration of DNA into the genome. A 150 kb cotton BAC DNA is reported to have [[ ]]been transferred [[ ]]into a specific *lox* site in tobacco by biolistic bombardment and *Cre-lox* site specific recombination.

Please amend the specification on page 94 line 16 to page 95 line 4 as follows:

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate media comprising the

nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as ~~WordPerfect~~ WORDPERFECT® and ~~Microsoft~~ MICROSOFT® Word, or represented in the form of an ASCII file, stored in a database application, such as ~~DB2, Sybase, Oracle,~~ DB2®, SYBASE®, ORACLE®, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (*e.g.*, text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

Please amend the specification on page 95 lines 5 to 18 as follows:

By providing one or more [[of]] nucleotide sequences of the present invention, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al.*, *Comp. Chem. 17*:203-207 (1993), the entirety of which is herein incorporated by reference) search algorithms on a ~~Sybase~~ SYBASE® system can be used to identify open reading frames (ORFs) within the ~~genome~~ genomes that contain homology to ORFs

or proteins from other organisms. Such ORFs are protein-encoding fragments within the sequences of the present invention and are useful in producing commercially important proteins such as enzymes used in amino acid biosynthesis, metabolism, transcription, translation, RNA processing, nucleic acid and a protein degradation, protein modification, and DNA replication, restriction, modification, recombination, and repair.

Please amend the specification on page 101 line 26 to page 102 line 12 as follows:

Two basic methods can be used for DNA sequencing, the chain termination method of Sanger *et al.*, *Proc. Natl. Acad. Sci. USA 74*:5463-5467 (1977), the entirety of which is herein incorporated by reference and the chemical degradation method of Maxam and Gilbert, *Proc. Natl. Acad. Sci. USA 74*:560-564 (1977), the entirety of which is herein incorporated by reference. Automation and advances in technology such as the replacement of radioisotopes with fluorescence-based sequencing have reduced the effort required to sequence DNA (Craxton, *Methods*, *2*:20-26 (1991), the entirety of which is herein incorporated by reference; Ju *et al.*, *Proc. Natl. Acad. Sci. USA 92*:4347-4351 (1995), the entirety of which is herein incorporated by reference; Tabor and Richardson, *Proc. Natl. Acad. Sci. USA 92*:6339-6343 (1995), the entirety of which is herein incorporated by reference). Automated sequencers are available from, for example, Pharmacia Biotech, Inc., Piscataway, New Jersey (Pharmacia ALF™), LI-COR, Inc., Lincoln, Nebraska (LI-COR® 4,000) and Millipore, Bedford, Massachusetts (Millipore BaseStation).

Please amend the specification on page 103 lines 5 to 14 as follows:

PHRED is used to call the bases from the sequence trace files ~~(http://www.mbt.washington.edu)~~ (www-mbt.washington.edu). ~~Phred~~ PHRED uses Fourier methods to examine the four base traces in the region surrounding each point in the data set in order to predict a series of evenly spaced predicted locations. That is, it determines where the peaks would be centered if there were no compressions, dropouts, or other factors shifting the peaks from their "true" locations. Next, PHRED examines each trace to find the centers of the actual or observed peaks and the areas of these peaks relative to their neighbors. The peaks are detected independently along each of the four traces so many peaks overlap. A dynamic programming algorithm is used to match the observed peaks detected in the second step with the predicted peak locations found in the first step.

Please amend the specification on page 104 lines 4 to 23 as follows:

This example illustrates the identification of combigenes within the rice genomic contig library as assembled in Example 2. The genes and partial genes that are embedded in such contigs are identified through a series of informatic analyses. The tools to define genes fall into two categories: homology-based and predictive-based methods. Homology-based searches (*e.g.*, GAP2, BLASTX supplemented by NAP and TBLASTX) detect conserved sequences during comparisons of DNA sequences or hypothetically translated protein sequences to public and/or proprietary DNA and protein databases. Existence of an *Oryza sativa* gene is inferred if significant sequence similarity extends over the majority of the target gene. Since homology-based methods may overlook genes unique to *Oryza sativa*, for which homologous nucleic acid molecules have not yet been identified in databases, gene prediction programs are also used.

Predictive methods employed in the definition of the *Oryza sativa* genes ~~included~~ include the use of the ~~GenScan~~ GENSCAN^{TM} gene predictive software program which is available from Stanford University (*e.g.* at the web site ~~http://gnomic/stanford.edu/GENSCANW.html~~ gnomic/stanford.edu/GENSCANW.html, and the Genemark.hmm for Eukaryotes program from Gene Probe, Inc (Atlanta, GA) ~~http://www.geneprobe.net/index.htm~~ www-geneprobe.net/index.htm. ~~GenScan~~ GENSCAN^{TM}, in general terms, infers the presence and extent of a gene through a search for "gene-like" grammar. GeneMark.hmm searches a file containing DNA sequence data for genes. It employs a Hidden Markov Model algorithm with a species-specific inhomogeneous Markov model of gene-encoding regions of DNA.

Please amend the specification on page 104 line 24 to page 105 line 18 as follows:

The homology-based methods that are used to define the *Oryza sativa* gene set included GAP2, BLASTX supplemented by NAP and TBLASTX. For a description of BLASTX and TBLASTX *see* Coulson, *Trends in Biotechnology* [[12]] *12*:76-80 (1994) and Birren *et al.*, *Genome Analysis*, [[1]] *1*:543-559 (1997). GAP2 and NAP are part of the Analysis and Annotation Tool (AAT) for Finding Genes in Genomic Sequences which was developed by Xiaoqiu Huang at Michigan Tech University and is available at the web site ~~http://genome.cs.mtu.edu/~~ genome.cs.mtu.edu/. The AAT package includes two sets of programs, one set DPS/NAP (referred to as "NAP") for comparing the query sequence with a protein database, and the other set DDS/GAP2 (refered to as "GAP2") for comparing the query sequence with a cDNA database. Each set contains a fast database search program and a rigorous alignment program. The database search program identifies regions of the query sequence that

are similar to a database sequence. Then the alignment program constructs an optimal alignment for each region and the database sequence. The alignment program also reports the coordinates of exons in the query sequence. *See* Huang, *et al.*, [[ ]]*Genomics 46*: 37-45 (1997). The GAP2 program computes an optimal global alignment of a genomic sequence and a cDNA sequence without penalizing terminal gaps. A long gap in the cDNA sequence is given a constant penalty. The DNA-DNA alignment by GAP2 adjusts penalties to accommodate introns. The GAP2 program makes use of splice site ~~consensuses~~ consensus in alignment computation. GAP2 delivers the alignment in linear space, so long sequences can be aligned. *See* Huang, *Computer Applications in the Biosciences 10*:227-235 (1994). The GAP2 program aligns the *Oryza sativa* contigs with a library of 42,260 *Oryza sativa* cDNAs.

Please amend the specification on page 105 line 26 to page 106 line 2 as follows:

NAP takes a nucleotide sequence, translates it in three forward reading frames and three reverse complement reading frames, and then compares the six tralsnations against a protein sequence database (*e.g.* the non-redundant protein (*i.e.*, nr-aa database maintained by the National Center for Biotechnology Information as part of ~~GenBank~~ GENBANK™ and available at the web site: ~~http://www.ncbi.nlm.nih.gov)~~ www-ncbi.nlm.nih.gov)).

Please amend the specification on page 106 line 21 to page 107 line 6 as follows:

The second homology-based method used for gene discovery is BLASTX hits extended with the NAP software package. BLASTX is run with the *Oryza sativa* genomic contigs as queries against the ~~GenBank~~ GENBANK™ non-redundant protein data library identified as "nr-

aa". NAP is used to better align the amino acid sequences as compared to the genomic sequence. NAP extends the match in regions where BLASTX has identified high-scoring-pairs (HSPs), predicts introns, and then links the exons into a single ORF prediction. Experience suggests that NAP tends to mis-predict the first exon. The NAP parameters are:

gap extension penalty = 1

gap open penalty = 15

gap length for constant penalty = 25

min exon length (in aa) = 7

minimum total length of all exons in a gene (in nucleotide) = 200

homology > 40%

Please amend the specification on page 107 lines 7 to 8 as follows:

The NAP alignment score and ~~GenBank~~ GENBANK™ reference number for best match are reported for each contig for which there is a NAP hit.

Please amend the specification on page 107 lines 15 to 27 as follows:

The ~~GenScan~~ GENSCAN™ program is "trained" with *Arabidopsis thaliana* characteristics. Though better than the "off-the-shelf" version, the ~~GenScan~~ GENSCAN™ trained to identify *Oryza sativa* genes proved more proficient at predicting exons than predicting full-length genes. Predicting full-length genes is compromised by point mutations in the unfinished contigs, as well as by the short length of the contigs relative to the typical length of a gene. Due to the errors found in the full-length gene predictions by ~~GenScan~~ GENSCAN™, inclusion of ~~GenScan~~ GENSCAN™-predicted genes is limited to those genes and exons whose

probabilities are above a conservative probability threshold. The ~~GenScan~~ GENSCAN™ parameters are:

weighted mean ~~GenScan~~ GENSCAN™ P value > 0.4

mean ~~GenScan~~ GENSCAN™ T value > 0

mean ~~GenScan~~ GENSCAN™ Coding score > 50

length > 200 bp

minimum total length of all exons ina gene = 500

Please amend the specification on page 107 line 28 to page 108 line 2 as follows:

The weighted mean ~~GenScan~~ GENSCAN™ P value is a probability for correctly predicting ORFs or partial ORFs and is defined as the $(1/\Sigma\, l_i)(\Sigma\, l_i\, P_i)$, where "l" is the length of [[a]] an exon and "P" is the probability or correctness for the exon.

Please amend the specification on page 108 lines 6 to 19 as follows:

The gene predictions from these programs are stored in a database and then combigenes are derived from these predictions. A combigene is a cluster of putative genes which satisfy the following criteria:

1) All genes making up a single combigene are located on the same strand of a contig;

2) Maximum intron size of a valid gene is 4000bp;

3) Maximum distance between any two genes in the same combigene is 200bp, as measured by the bases between adjacent ending exons;

4) If an individual gene is predicted by NAP it has at least 40% sequence identity to its hit;

5) If an individual gene is predicted by GAP2 it has at least 92% sequence identity to its hit;

6) If an individual gene is predicted by ~~GenScan~~ GENSCAN™ the weighted average of the probabilities calculated for all of its exons is not less than 0.4. The gene boundaries of a ~~GenScan~~ GENSCAN™-predicted gene are determined while taking into account only exons.

Please amend the specification on page 109 lines 44 to 45 as follows:

Indicates the gene-predicting program used. These programs are ~~GenScan~~ GENSCAN™, AAT/NAP, AAT/GAP, TBLASTX or Genemark.hmm.

Please amend the specification on page 110 lines 24 to 29 as follows:

Each sequence in the ~~GenBank~~ GENBANK™ public database is arbitrarily assigned a unique NCBI gi (National Center for Biotechnology Information ~~GenBank~~ GENBANK™ Identifier) number. In this table, the NCBI gi number which is associated (in the same row) with a given contig or singleton refers to the particular ~~GenBank~~ GENBANK™ sequence which is the best match for that sequence. If the genomic sequence aligns to a cDNA from Monsanto's SeqDB, the name of the cDNA sequence is named.